Erschienen in: Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" / Komp'juternaja lingvistika i intellektual'nye tehnologii. Po materialam ezhegodnoj Mezhdunarodnoj konferencii „Dialog". Bekasovo, 26.-30. Mai 2010. Moskva: Izdatel'stvo Rossijskogo Gosudarstvennogo Gumanitarnogo Universiteta (RGGU) / Staatliche Universität für Sozial- und Geisteswissenschaften Moskau (RGGU) 2010, S. 622-628.

Using a domain ontology for the semantic-statistical classification of specialist hypertexts

Noah Bubenhofer (bubenhofer@ids-mannheim.de),
Roman Schneider (schneider@ids-mannheim.de)

Institute for German Language (IDS), Mannheim/Germany

In this feasibility study we aim at contributing at the practical use of domain ontologies for hypertext classification by introducing an algorithm generating potential keywords. The algorithm uses structural markup information and lemmatized word lists as well as a domain ontology on linguistics. We present the calculation and ranking of keyword candidates based on ontology relationships, word position, frequency information, and statistical significance as evidenced by log-likelihood tests. Finally, the results of our machine-driven classification are validated empirically against manually assigned keywords.

1. Introduction

Research into ontologies has received much attention for the last years [16] [17] [18]. Due to its practical use for common tasks related to knowledge sharing and publication, it has been subject of study in most different scientific communities. Ontologies are often seen as enabling technology for information sharing, with their ability to be easily reused being a key factor for successful application scenarios [4] [6] [8] [15]. On the web, which represents a large universe of mostly unclassified semi-structured hypertexts, semantic techniques and technologies open up new strategies for information retrieval and text classification [5].

The Institute for German Language (IDS) in Mannheim is the central institution for research and documentation of the German language. It hosts several specialist resources, including the hypertextual information systems Grammis and ProGr@mm and a terminological ontology [12] [13] [14]. Since only less than 40 % of the hypertexts are classified with manually assigned keywords, our goal is to gain insight of how ontology features can affect automatic semantic-statistical classification. We introduce the resources as far as necessary to understand our test-bed, and then present a self-conducted empirical case study to verify the feasibility of our approach.

2. Hypertext resources

Grammis is a specialist hypertext resource that brings together terminological, lexicographical, and bibliographical information about German grammar. Initiated more than a decade ago, it combines traditional description of grammatical structures with the results of corpus-based studies and hypermedia design principles. Considering that the grammar of human languages is a highly complex scientific domain, the project authors use hypertext chunking and linking as well as multimedia extensions like spoken language excerpts and graphical explanations in order to address a broad target audience with heterogeneous foreknowledge. Their goal is to present a comprehensive overall picture of contemporary German grammar from a syntactic, semantic, and functional perspective. Today, Grammis[1] is the most prominent academic information system dedicated to German Grammar, with consistently more than 50,000 page impressions per month. ProGr@mm[2] is an e-learning system for schools, colleges, and uni-

[1] Short for: Grammatical Information System (*http://www.ids-mannheim.de/grammis/)*. The authors of this paper are members of the Grammis project team.

[2] Short for: Propaedeutic Grammar (*http://www.ids-mannheim.de/progr@mm/*)

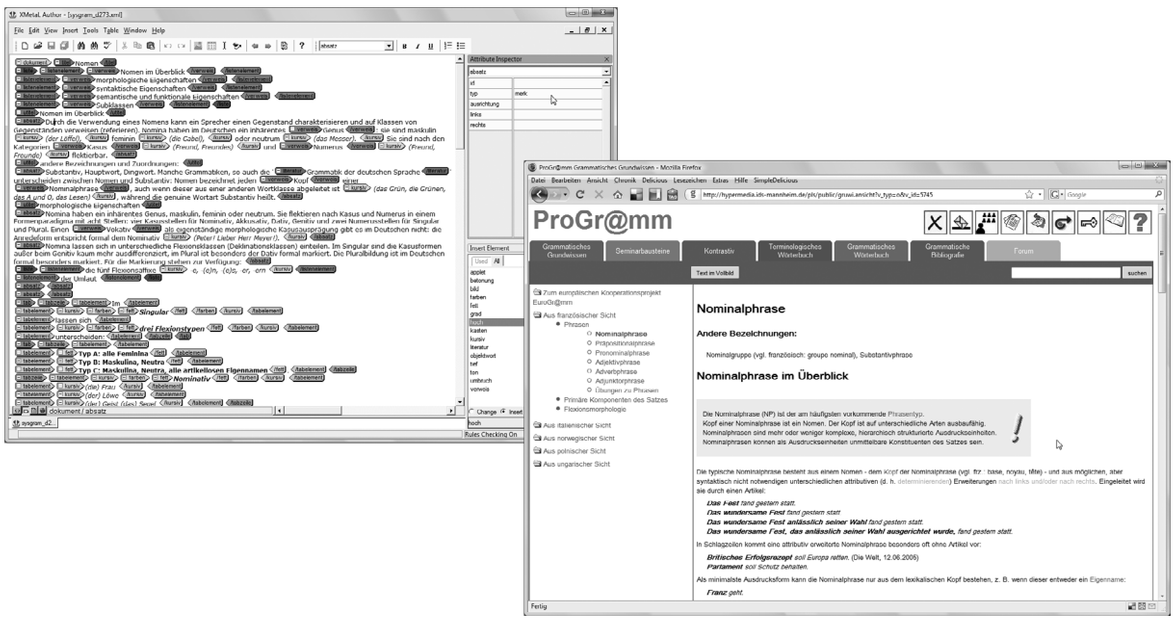

Fig. 1: XML stored inside the database (left) and converted to HTML (right)

versities, and didactically prepared for online learning. A special module covers selected grammatical topics from the perspective of other European languages and is well-suited for students and teachers of German as a foreign language. Functional add-ons are guided tours, personal notes, and discussion forums for the educational community.

From a technical point of view, both Grammis and ProGr@mm can be described as XML- and database-driven web information systems, whose semi-structured hypertext nodes (instances) conform to the Grammis Markup Language (GrammisML). GrammisML defines detailed constraints on the instance's logical structure, allowing for subsequent cross-media publishing ("one source fits all"). It provides conventional block elements like paragraphs, lists, or tables, as well as specific markup structures for the coding of grammatical metadata, typed hyperlinks, etc. Using a web-based authoring frontend, arbitrary keywords and object words/phrases for retrieval operations can be assigned manually. Parsing, analysis, and transformation of the hypertext resources are conducted using established technology like XPath, XQuery, XSQL, and XSL Transformation [11].

3. The domain ontology

Not just since the proclamation of the Semantic Web [2], semantic resources are among the most prominent add-ons and tools for information retrieval. Domain ontologies, organizing specialist terms (concepts) and their interconnections (relationships), can make a most valuable contribution to the analysis, classification, and finding of documents on the web — not least in the context of academic publications [3]. This is due to the both simple and unfortunate fact that scientific terminology is often far from being consistent. Especially in the field of linguistics, different theories, schools of thought, or even authors not only name things differently, but even assign varying meanings to identical terms. A semantically enriched retrieval application for the exploration of linguistic resources should incorporate these theory-related details so that it can offer appropriate solutions. As a consequence, we integrated a domain ontology for linguistic/grammatical terminology. The semiautomatic detection of concepts as well as the modeling of relationships has been conducted using statistical methods on large general language corpora and specialist language corpora.[3] Broadly speaking, in order to bring together theoretical desiderata with practical demands and limitations, we combine well-established standards of ontological engineering — e. g., the use of ISO-2788/ANSI Z39.19 compliant hyponymy/meronymy relationship types like Broader Term Generic (BTG) or Broader Term Partitive (BTP) — with terminological modeling principles — e. g., termsets, expanded by theory-related attributes and explicit linking of individual concepts belonging to different Termsets [1].

Figure 2 illustrates our ontology model. It covers three termsets, indicated by dotted border lines. The bottom termset contains the two concepts *Verbgruppe* and *Verbalphrase*, recognizable by rectangles with rounded corners. *Verbgruppe* is characterized by a theory-related attribute named *IDS*', meaning that it is used primarily

[3] See [12] for a description of the ontology building process; [14] describes the ontology in greater detail.

when referring to the *IDS Grammar of German Language*. The concept *Verbalphrase* consists of four lexical entries:

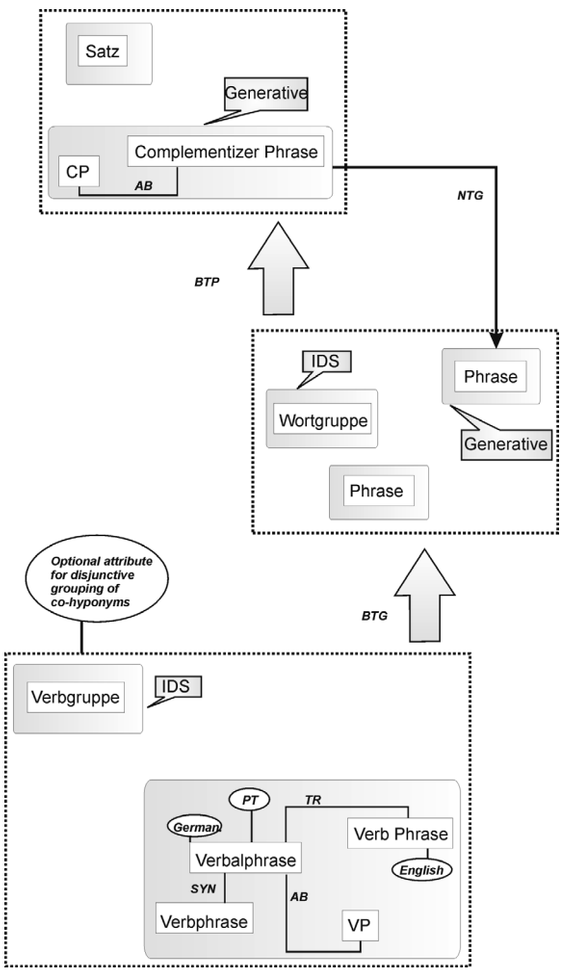

Fig. 2: Grammis ontology modeling structure

- *Verbalphrase* with a marker for Preferred Term (PT) and a language attribute (*German*)
- *Verbphrase* linked to the former by a synonymy relation (SYN)
- *VP* linked by a abbreviation relation (AB)
- *Verb Phrase* with language attribute (*English*) and translation relation (TR)

The termset is linked with its hyperonym by a Broader Term Generic (BTG) relation. In order to clarify the benefit of linking not only termsets, but also concepts, our example illustrates the relationships between *Phrase* (engl. *phrase*) and *Satz* (engl. *sentence*). Basically, the corresponding termsets are connected with the help of a Broader Term Partitive (BTP) relation (meronymy). Beyond this, since generative grammars usually classify sentences as phrases (*complementizer phrases*), only these two concepts — singled out by a theory-related attribute — are linked by a Narrower Term Generic (NTG) relation (hyponymy). This should facilitate communication between people or computers using different terminological systems.

4. The classification process

The goal of the classification process is to find terms (keywords) describing the content of a hypertext. We use the following information for our algorithm:

- The hypertexts contain XML-coded markup like paragraphs, lists, tables etc., but also specific grammatical metadata and links to the grammatical dictionary.
- For the classification process the hypertexts were lemmatized and part-of-speech tagged using the "TreeTagger" [10] and a training file for German.
- The source for possible keywords is our ontology, that can be accessed by functions such as „get hypernyms of a term *x* up to *n* levels" or „get synonyms of term *x*" etc.

4.1. The ranking algorithm

For the classification process we stored the hypertexts as a lemmatized word list which also contains the type of the paragraph the word is used in (title, subtitle or definition). We omitted words that are used in examples and tables: Examples contain object language that should not be used as a source for keyword candidates. Tables also often list object language and contain word chunks or fragments, because they are used for the presentation of inflection paradigms and the like. The basic idea of our classification algorithm is the following:

1. We select for each text all the terms that are also part of the ontology.
2. For each term, we assume broader terms one level above as additional keyword.
3. For each term, a rank is calculated that reflects its importance within the text. We use basically three factors to calculate importance:
   a. Frequency: More frequent terms are more important than rare ones.
   b. Position: If a term is mentioned in a title, subtitle or a definition or is used as a link to the grammatical dictionary, then it is supposed to be more important.
   c. Statistical significance: The relative frequency compared to the mean frequency of the term in all the other texts is calculated using a log-likelihood test.

These three factors are combined to an overall score. Frequency and position are calculated by counting the occurrences of the term in question multiplied by a weight depending on its position. In our standard procedure we used: titles = 6, subtitles = 4, definitions and „Merksätze" = 2, all other positions = 1. The statistical significance is calculated using the log-likelihood test [9]:

$$LL = 2*((a*\log (a/E_1)) + (b*\log (b/E_2)))$$

In this formula, a and b are the raw frequencies of the term in the text and the whole corpus respectively. E_1 and E_2 are the expected frequencies in the text and the whole corpus. The calculated value expresses the difference of the relative frequency to the total corpus. The higher the value, the higher is the significance of the term for the specific text. The following example demonstrates the difference between raw frequencies and relative frequency: Table 1 shows the frequencies and significance values of a hypertext node on valency ("Valenz").[4]

The keywords are ordered by their significance for the text (column „LLR"). Column „frequency" contains the raw frequencies, and „weighted frequency" stands for the frequencies weighted by the position in the text. The list also contains terms that are not mentioned literally, but are broader terms of a token („source term"). The most frequent term is *Valenzträger*, but according to the raw and the weighted frequency, *Valenzträger* would be on a lower rank. And vice versa: A very often used term like *Adjektiv* is not significant enough for a text on (verb) valency to rank in a top position ordered by significance.

[4] *http://hypermedia.ids-mannheim.de/pls/public/sysgram.ansicht?v_typ=d&v_id=2871*

4.2. The final ranking

The algorithm produces two different rankings: One ranking reflects the combination of frequency of the term and its position, the other ranking represents the significance of the term. Both aspects influence the final ranking. We combined the two rankings in the following way: The rank is transformed to a score by inverting the rank position. We then sum up the two scores and get a final ranking. In addition, we omit keywords with raw frequency 1 which tend to get very high LLR values but are not important enough to be included into the keyword list. When applying the algorithm to the valency hypertext node (see table 1 above for the raw frequencies), we get the final ranking as shown in table 2.

The number of keyword candidates depends on how congruent the two lists of the highest ranked terms are. Table 2 is based upon the combination of two top 10 lists and both lists contain more or less the same terms in different order. Therefore the merged list contains only two terms more than the two source lists. Intuitively, table 2 satisfyingly reflects the text about valency, similar to other hypertext nodes we evaluated manually. But, as described in the following section, we tried to further evaluate the lists for better results.

Table 1: Comparison of different measures for the frequency of terms in the valency hypertext node

ID	Type	keyword Candidate	Frequency	Weighted frequency	LLR	Source term
2871	d	Valenzträger	8	17	70,93	Valenzträger
2871	d	syntagmatische Beziehung	11	23	64,89	Valenz
2871	d	Valenz	11	23	64,89	Valenz
2871	d	Komplement	18	18	54,44	Komplement
2871	d	Leerstelle	3	3	26,21	Leerstelle
2871	d	Wortart	15	33	13,68	Verb
2871	d	Verb	15	18	13,68	Verb
2871	d	Nominalgruppe	2	2	13,33	Nominalgruppe
2871	d	Modifikator	10	14	10,94	Adjektiv
2871	d	Adjektiv	10	13	10,94	Adjektiv
2871	d	Satzadverbial	10	13	10,94	Adjektiv
2871	d	Nomen	10	13	9,85	Nomen
2871	d	Bedeutung	6	6	9,66	Bedeutung
2871	d	Verbvalenz	1	1	6,25	Verbvalenz
2871	d	Eigenschaft	4	4	4,58	Eigenschaft
2871	d	Prädikat	4	4	4,58	Eigenschaft
2871	d	Form	6	6	3,51	Form
2871	d	Ergänzung	1	1	3,34	Ergänzung
2871	d	Infinitivkonstruktion	1	1	3,15	Infinitivkonstruktion
2871	d	Anhebung	1	1	3,15	Infinitivkonstruktion
2871	d	Infinitkonstruktion	1	1	3,15	Infinitivkonstruktion
2871	d	Nominalphrase	4	4	2,65	Nominalphrase

Table 2: Final ranking of the terms in the text „Verbvalenz“

ID	Type	Keyword Candidate	Score
2871	d	Valenz	17
2871	d	syntagmatische Beziehung	17
2871	d	Wortart	15
2871	d	Valenzträger	14
2871	d	Komplement	12
2871	d	Verb	10
2871	d	Leerstelle	6
2871	d	Modifikator	5
2871	d	Nominalgruppe	3
2871	d	Satzadverbial	2
2871	d	Adjektiv	2

5. Evaluation of the classification results

5.1. Evaluation results

Some of the hypertext nodes are already classified by manually assigned keywords, using an uncontrolled vocabulary. These keywords are a measure to evaluate our automated classification and to experiment with different settings of the classification algorithm. Currently, the algorithm cannot cope with multi-word units, therefore we only analyze texts with one or more single-word keywords. Table 3 shows how the change of some parameters of the classification algorithm — e. g., weight of position (title, subtitle, etc.) — affects the matching of manually and automatically assigned keywords. We differentiate three matching levels: level 1 counts documents, that at minimum have one correspondence of manually and automatically assigned keywords. At level 2 at least 50 %, and at level 3 all of the manual keywords need to be matched by the automatic ones.

Table 3: Evaluation of the automatic assigned keywords

Version	Parameters	Level 1: Matching documents Min. 1 KW	Level 2: Min. 50 % KW	Level 3: Min. 100 % KW
1	*Default version* Weight of positions: titel = 10, subtitle = 4, definitions and „Merksätze“ = 2 Source lists: top 10	79,34 % 657/828	37,68 % 312/828	22,4 % 186/828
2	*More keywords* Equal to default version, but: Source lists: top 20	83,69 % 693/828	48,18 % 399/828	29,71 % 246/828
3	*More keywords* Equal to default version, but: Source lists: top 40	85,02 % 704/828	52,29 % 433/828	32,97 % 273/828
4	*More keywords* Equal to default version, but: Source lists: top 100	85,02 % 704/828	52,54 % 435/828	33,33 % 276/828
5	*Titles version* Equal to default version, but: titel = 30, subtitle = 10	79,59 % 659/828	38,53 % 319/828	23,19 % 192/828
6	*Versions with more keywords lead to the same results than versions 2–4 above*			
7	*No hypernyms* Equal to default version, but: Only literally used words are keyword candidates, no hypernyms.	79,10 % 655/828	37,07 % 307/828	22,46 % 186/828
8	*More hypernyms* Equal to default version, but: Hypernyms up to 2 levels above in the hierarchy	78,02 % 646/828	37,44 % 310/828	21,98 % 182/828
9	*More keywords* Equal to “more hypernyms” version (8), but: Source list: top 100	85,51 % 708/828	53,5 % 443/828	34,54 % 286/828

The evaluation illustrates two key issues for successful keyword detection:

- Getting all possible keyword candidates out of the text (tested with versions 1, 5, 7 and 8 in table 3).
- Putting the keyword candidates into the right ranking order, so that the top 10 ranking reflects the text content (tested with versions 2–4, 6 and 9 in table 3).

The first evaluation results are not too impressive: A 50 % matching of automatically and manually assigned keywords is only achieved at about 37 % of the documents (table 3, 1). About 80 % of the documents have at minimum one correspondence. Crucial seems the number of keywords that are included into the final list of keyword candidates. If this number is being increased, the matching scores also get better (table 3, 2–4). But even if the source lists contain 100 keyword candidates, only 52 % of the documents have matches at a 50 % level (85 % at level 1). If other parameters are changed, the score does not increase significantly: Neither accepting less nor more hypernyms (table 3, 7–8) has a substantial impact on the matching score. Only a higher weight of title positions (table 3, 5) slightly increases the score.

5.2. Discussion

These results interfere with our first impression when we intuitively evaluated documents without any manual keywords. Therefore, the manual classification process has to be examined. In 262 (32 %) of 828 documents, at minimum 80 % of all manually chosen keywords are not used at all within the hypertext nodes, even if the most narrow terms are taken into consideration. The reasons for that are manifold:

- Tagging issues influence the matching results: The TreeTagger does not lemmatize some plural forms (e. g., *Pronomina*) correctly. This leads to a mismatch in hypertext nodes where only the plural form is used.
- The fact that at the moment we cannot cope with multi-word units also affects the evaluation of the manual classification process.
- Our human classifiers tend to choose keywords that are neither mentioned in the hypertext node nor are close hypernyms of text words.

The above mentioned hypertext node ("Relativ-Elemente") also shows that different keynote annotators could disagree on the best solution (bad inter-rater reliability). *Pronomen* and *Wortart* are the manually assigned keywords, but another rater perhaps would also or instead set *Relativsatz*, *Relativ-Element* (as used in the title of the text) or something else as a keyword. Table 4 shows the automatically assigned keywords to the text.

Table 4: Final ranking of the terms in the text „Relativ-Elemente"

ID	Type	Keyword Candidate	Score
368	d	Phrase	11
368	d	grammatische Kategorie	10
368	d	Relativsatz	10
368	d	eingeleiteter Nebensatz	9
368	d	Einheitenkategorie	8
368	d	Relativ-Element	8
368	d	nicht-verbaler Ausdruck	7
368	d	Pronominalphrase	7
368	d	Einbettung	7
368	d	Proposition	6
368	d	Verkettungsverfahren	6
368	d	restriktiv	5
368	d	Präpositionalphrase	4
368	d	semantische Relation	4
368	d	phrasale Kategorie	2
368	d	Nominalphrase	1

6. Conclusion

The discussion shows the demand for a gold standard regarding the automatic detection of keywords for specialist texts. But the establishing of such standards seems difficult due to the fact that different (hypertext) publications even today mostly use different microstructures. An orientation to existing guidelines like TEI would possibly ease the determination of default settings for position parameters like title, subtitle, paragraph types, etc. Beyond that, controlled vocabularies for the manually assigned keywords — or, alternatively, the integration of user-independent data like social bookmark tags or folksonomies — would surely affect the congruity with machine-detected terms. Nevertheless, the first results of our ontology-based approach encourage for further application in the context of information retrieval and classification — and for methodological comparisons with other approaches for automatic keyword extraction.

References

1. *Beißwenger, M.* TermNet — ein terminologisches Wortnetz im Stile des Princeton WordNet // TU Dortmund: Institut für deutsche Sprache und Literatur, 2008.
2. *Berners-Lee, T. / Hendler, J. A. / Lassila, O.* The Semantic Web // Scientific American 284 (5), 2001. P. 34–43.
3. *Chiarcos, C.* An Ontology of Linguistic Annotations // LDV-Forum/Journal for Computational Linguistics and Language Technology, 2008. 23 (1). P. 1–6.
4. *Fensel, D.* Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce // New York: Springer, 2001.
5. *Gottron, T. / Schneider, R.* A Hybrid Approach to Statistical and Semantical Analysis of Web Documents // Merabti, M. (Ed.): Proceedings of The Fifth IASTED European Conference on Internet and Multimedia Systems and Applications (EuroIMSA). Acta Press, 2009. P. 115–120.
6. *Gruber, T. R.* Ontology of Folksonomy: A Mash-up of Apples and Oranges // International Journal on Semantic Web & Information Systems, 2007. Vol. 3.2. P. 1–11.
7. *Gruber, T. R.* Ontology // Liu, L. / Tamer Özsu, M. (Eds.): Encyclopedia of Database Systems. New York: Springer, 2009.
8. *Maedche, A.* Ontology Learning for the Semantic Web // Kluwer Academic Publishers, 2002.
9. *Rayson, P. / Garside, R.* Comparing corpora using frequency profiling // Proceedings of the workshop on Comparing Corpora, 38th annual meeting of the Association for Computational Linguistics (ACL), Hong Kong, 2000. P. 1–6.
10. *Schmid, H.* Probabilistic Part-of-Speech Tagging Using Decision Trees // Proceedings of the International Conference on New Methods in Language Processing, Manchester, 1994. P. 44–49.
11. *Schneider, R.* E-VALBU: Advanced SQL/XML processing of dictionary data using an object-relational XML database // SDV — Sprache und Datenverarbeitung/International Journal for Language Data Processing, 2008. Vol. 32.1/2008. P. 33–44.
12. *Schneider, R.* A Database-driven Ontology for German Grammar // Rehm, G. / Witt, A. / Lemnitzer, L. (Eds.): Data Structures for Linguistic Resources and Applications. Proceedings of the Biennial GLDV Conference 2007. Narr, 2007. P. 305–314.
13. *Schneider, R.* Benutzeradaptive Systeme im Internet: Informieren und Lernen mit GRAMMIS und ProGr@mm // Mannheim: Institut für Deutsche Sprache (IDS), 2004.
14. *Sejane, I.* Database-Driven Access to Heterogeneous XML-Contents Using Domain Ontology of German Grammar // SDV — Sprache und Datenverarbeitung/International Journal for Language Data Processing, 2008. Vol. 32.1/2008. P. 71–87.
15. *Simperl, E.* Reusing ontologies on the Semantic Web: A feasibility study // Data & Knowledge Engineering, 2009. Vol. 68/10. P. 905–925.
16. *Staab, S. / Studer, R.* Handbook on Ontologies. International Handbooks on Information Systems // New York: Springer, 2009.
17. *Studer, R. / Davies, J. / Warren, P.* Semantic Web Technologies — Trends and Research in Ontology-Based Systems // John Wiley & Sons, 2006.
18. *Tran, T. / Cimiano, P. / Rudolph, S. / Studer, R.* Ontolgy-Based Interpretation of Keywords for Semantic Search // Proceedings of the 6th International Semantic Web Conference, 2007. P. 523–536.